Compressed Domain Automatic Level Control based on ITU-T G.722.2

Yu-wen HE, Chang-chun BAO and Bing-yin XIA
Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering
Beijing University of Technology, Beijing, 100124, China
E-mail: iamhyw@emails.bjut.edu.cn, chchbao@bjut.edu.cn, xby-abc@emails.bjut.edu.cn

***Abstract*—In order to control the speech level automatically in the 3G mobile networks where Transcoder Free Operation (TrFO) is adopted, a compressed domain automatic level control (ALC) method is proposed based on ITU-T G.722.2 speech codec. The level of decoded speech is first measured by P.56 tool. Then, based on the difference between the measured level and the target level, the gain of ALC is calculated. Finally, the codebook gain parameters of speech codec are modified according to the ALC gain and the level of decoded speech is adjusted to the target level. The result of performance evaluation shows that, in comparison with the traditional method, in which the speech signal is first decoded, then level adjusted by ITU-T P.56 tool, and finally re-encoded, the proposed method saves 65% to 75% of the computational complexity. The difference between the output level and the target level is smaller and the objective speech quality is much better.**

Keywords—speech coding; automatic level control; compressed domain; G.722.2; P.56

I. INTRODUCTION

Due to the change of voice level of speakers, the different call environments and some other issues of speech transmission in the mobile communications, there are always fluctuations, even overloads, in the speech amplitude, which makes the speech quality degraded seriously. Automatic Level Control (ALC) can solve the problem adaptively by amplifying weak speech or attenuating strong speech [1]. Therefore, ALC has been widely studied in the last decades.

In the conventional approaches of ALC, the speech amplitude is scaled directly in linear domain. In [2] - [4], with the assistance of Voice Activity Detection (VAD), ALC is realized by adjusting the difference between the power of active speech and the target level. On the other hand, in ITU-T recommendation P.56 [5], the speech level is measured based on the statistic distributions of the speech amplitudes of the whole sentence in the quantization intervals. The difference between the measured level and the target level is calculated to control the level off-line. But in the 3G network using the TrFO technology, the speech signal is transmitted in the form of encoded bit-stream between the two mobile terminals, even in the core network. If linear domain ALC methods are applied in this kind of networks, the speech should first be decoded, then level controlled, and finally re-encoded. In the practical applications, the additional computational complexity, delay, and the degradation of speech quality are usually not acceptable. Therefore, the linear domain approaches are not suitable for this kind of networks.

In order to solve this problem, the ALC methods which control the speech level through the modification of codec parameters, i.e., the compressed domain ALC method, have been paid more and more attentions. In [6], a dynamic scaling method of encoded speech in the compressed domain is proposed. The codebook gain parameters of the bit-stream are modified to control speech level depending on the scale factor. In this method, the scale factor is set manually in advance, which makes it not feasible for the real-world applications.

In this paper, based on the approach of P.56 and the method in [6], we propose a compressed domain automatic level control method based on ITU-T G.722.2 [7]. The proposed method measures the real-time level of input decoded speech by the P.56 tool, and the beginning of ALC operation is determined by the stationarity of real-time level. The scale factor is computed adaptively using the target level and the real-time level. Based on the scaling factor, the adaptive and fixed codebook gains are adjusted and joint-quantized to modify the corresponding part of the bit-stream.

This paper is organized as follows. The basic principle of ITU-T G.722.2 codec is reviewed in Section II. Section III describes the proposed method. Finally, the performance evaluation is presented in Section IV, and Section V gives the conclusions.

II. REVIEW OF ITU-T G.722.2 CODEC

ITU-T G.722.2 is a wideband speech codec based on CELP model, which is also called Adaptive Multi-Rate Wideband (AMR-WB). The sampling rate of input speech is 16 kHz with the bandwidth of 7 kHz. The codec provides 9 bit rates ranging from 6.6 kbps to 23.85 kbps. It is appropriate for the wired and wireless communication networks, such as ISDN, the PSTN, VoIP, WCDMA and 3G networks, etc.

In the CELP model, the speech signal $s(n)$ can be expressed as the convolution of excitation signal $u(n)$ and synthesis filter's impulse response $h(n)$:

$$s(n) = u(n) * h(n) \tag{1}$$

where n is the sample index, $u(n)$ is given by:

978-1-4673-5604-6/12/$31.00 ©2012 IEEE

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2012	**N/A**	**-**

4. TITLE AND SUBTITLE
Compressed Domain Automatic Level Control based on ITU-T G.722.2

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering Beijing University of Technology, Beijing, 100124, China

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADA587934 . IEEE International Symposium on Signal Processing and Information Technology (12th) (ISSPIT 2012) Held in Ho Chi Minh City, Vietnam on December 12-15, 2012 . AOARD - CSP-131010.

14. ABSTRACT
In order to control the speech level automatically in the 3G mobile networks where Transcoder Free Operation (TrFO) is adopted, a compressed domain automatic level control (ALC) method is proposed based on ITU-T G.722.2 speech codec. The level of decoded speech is first measured by P.56 tool. Then, based on the difference between the measured level and the target level, the gain of ALC is calculated. Finally, the codebook gain parameters of speech codec are modified according to the ALC gain and the level of decoded speech is adjusted to the target level. The result of performance evaluation shows that, in comparison with the traditional method, in which the speech signal is first decoded, then level adjusted by ITU-T P.56 tool, and finally re-encoded, the proposed method saves 65% to 75% of the computational complexity. The difference between the output level and the target level is smaller and the objective speech quality is much better.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**6**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

$$u(n) = g_c(m)c(n) + g_p(m)v(n) \tag{2}$$

where $g_p(m)$ and $g_c(m)$ are the adaptive and fixed codebook gains, respectively. $v(n)$ and $c(n)$ are the adaptive and fixed codebook vectors, respectively. m is the sub-frame index.

The adaptive codebook excitation models the periodicity of the speech signal, it can be written as:

$$v(n) = u(n-T) \tag{3}$$

where T is the pitch delay.

It is known that, speech level is proportional to the instantaneous power of speech signal. That is to say, the control of speech level can be achieved by scaling the excitation signal with synthesis filter unchanged. According to (2), $g_c(m)$ and $g_p(m)$ represent the amplitude information of excitation signal. As shown in (3), the adaptive and fixed codebook excitation signals are related to each other through the process of long-term prediction, modifying only one of the gain parameters may have negative effect on the speech quality. For this reason, the gain parameters should be jointly modified [6] to constrain the loss of speech quality in an acceptable range during ALC operations.

III. Compressed Domain Automatic Level Control

The proposed compressed domain ALC method is carried out on the encoded bit-stream of ITU-T G.722.2 codec. The block diagram is shown in Fig. 1.

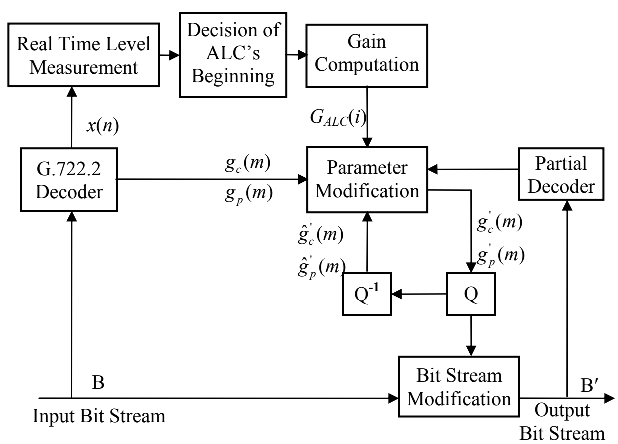


Figure 1. The block diagram of ALC in the compressed domain

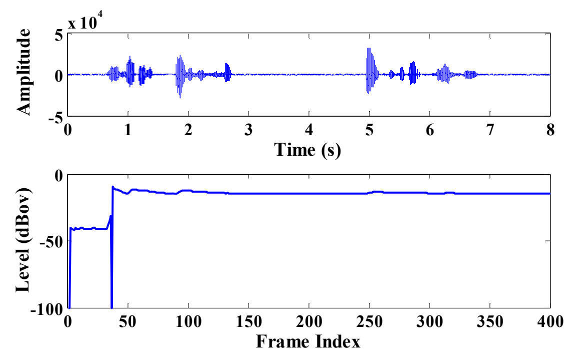
First, the input bit-stream B is decoded to get the time domain signal $x(n)$ by G.722.2 decoder, and the real-time level of $x(n)$ is measured by the P.56 tool. Then, the ALC operation will begin when the real-time level becomes stationary, and the ALC gain $G_{ALC}(i)$ (i is the frame index) is computed according to the difference between the real-time and target level. Next, the codebook gain parameters, $g_p(m)$ and $g_c(m)$, are jointly modified depending on $G_{ALC}(i)$. Finally, the modified adaptive and fixed codebook gains, $g'_p(m)$ and $g'_c(m)$, are re-quantized and written back to the corresponding part of the bit-stream. The output bit stream B′ is decoded partially in order to get the excitation signal for ALC operation in the next sub-frame.

In Fig. 1, $\hat{g}'_p(m)$ and $\hat{g}'_c(m)$ are the quantized adaptive and fixed codebook gains, respectively. The module Q is the gain quantization and Q^{-1} represents the inverse quantization. The details of the proposed method will be described in the following sub-sections.

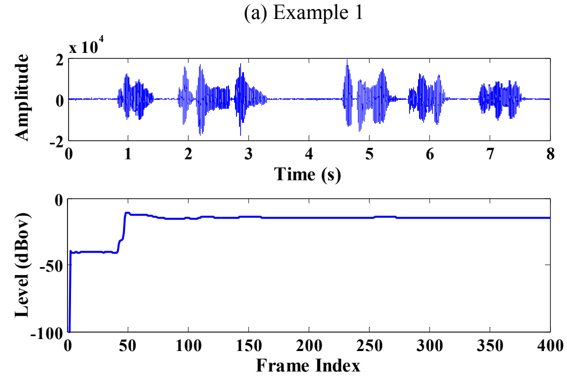
A. Real Time Speech Level Measurement

Method B of ITU-T P.56 [5] is the standardized method for the offline objective measurement of speech level, and the output active speech level is a quantity proportional to instantaneous power over the aggregate of time during which the speech is present. In order to meet the requirement of real-time processing, we extract the active speech level of each frame estimated by the P.56 tool as an approximate estimation of the real-time speech level.

Fig. 2 shows the waveforms of two noisy speech samples and the corresponding real-time speech level measured by the P.56 tool. The noisy speech signals with the SNR of 18dB are obtained by mixing the clean speech of -15dBm0 with car noise. Here, dBm0 and dBov are the two most commonly used level units. The dBm0 is often used in the digital transmission networks for different types of signals (speech, modem, fax, etc.). While in the digital signal processing equipments, such as speech codec, the speech level unit is dBov. For A-law PCM format, the relationship between overload (dBov) and maximum levels (dBm0) is dBm0 = dBov + 6.15 [8].



(a) Example 1



(b) Example 2

Figure 2. Examples of the real-time level for two noisy speech samples

B. Decision of ALC's Beginning

The real-time level has many fluctuations in the initial segment of speech material. After active speech appears, more statistical information about speech amplitude is accumulated such that the real-time level gets stationary gradually. In order to get satisfactory performance, we should determine when the ALC operation begins.

Different decision methods are proposed for the two types of variations in real-time level (as shown in Fig. 2). On one hand, as shown in Fig. 2 (a), when the real-time level jumps to the initial level of -100dBov for twice, ALC process will begin and the flag of ALC is set to one. On the other hand, when the real-time level approaches the actual level gradually without large fluctuations as shown in Fig. 2 (b), the level stability factor R_{AL} is proposed to determine whether the real-time level is stationary. It is defined as the ratio of the real-time level of current frame and its long-term average in the previous L_{AL} frames, which can be expressed as:

$$R_{AL}(i) = \frac{AL(i)}{\frac{1}{L_{AL}}\sum_{j=1}^{L_{AL}} AL(i-j)}, \quad AL \neq -100dBov \tag{4}$$

where i is the frame index, $AL(.)$ is the active speech level. The condition, $AL \neq -100dBov$, excludes the silence period in which the active level is set to the initial value by P.56 tool, and it will make the average value inaccurate. Fig. 3 depicts the level stability factor R_{AL} and the real-time speech level, The R_{AL} of the first L_{AL} frames is initialized to zero.

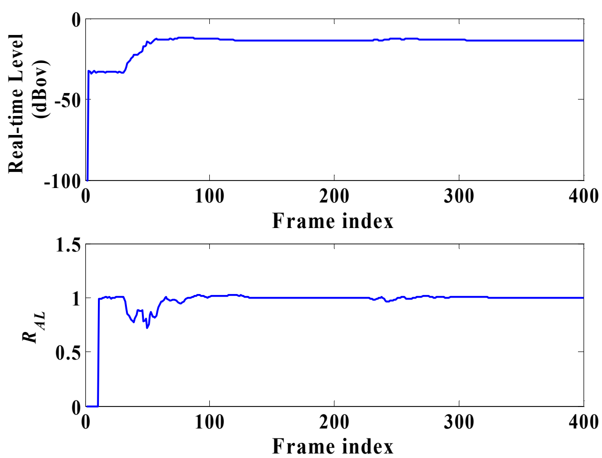


Figure 3. The real-time level and real-time level stability factor

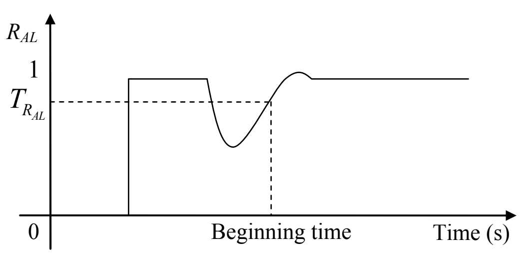


Figure 4. The real-time level stability and its threshold

Fig. 3 shows that the real-time level stability factor R_{AL} has an apparent fluctuation around 1 when the real-time level gets closer to the actual level.

The variation of R_{AL} in the initial segment of speech material is shown in Fig. 4. $T_{R_{AL}}$ is the threshold for the stability factor, its value is empirically determined in the range of 0 to 1. If R_{AL} has crossed the threshold for three times from the start of the speech, it can be determined that the real-time level is stationary and active speech has appeared. Then the flag of ALC is set to one and ALC process begins. Whether R_{AL} crosses the threshold or not is determined by:

$$(R_{AL}(i) - T_{R_{AL}})(R_{AL}(i-1) - T_{R_{AL}}) < 0 \tag{5}$$

where i is the frame index. From Fig. 4, it can be inferred that, $T_{R_{AL}}$ is closely related to the convergence time of ALC algorithm. With smaller threshold value, the ALC operation will start earlier. Considering both the convergence time and the applicability in variant conditions, the threshold is set to 0.95 in this paper.

C. Gain Computation

The difference between the target level and the real-time level in decibels can be converted to the scaling factor of the excitation signal, namely the ALC gain, by the following relationship:

$$G_{ALC}(i) = 10^{(AL_{dest} - AL(i))/20} \tag{6}$$

where AL_{dest} and $AL(i)$ are the target level and the input level in dBov, respectively.

D. Parameter Modification

The adaptive and fixed codebook gains, $g_p(m)$ and $g_c(m)$, are modified according to the ALC gain $G_{ALC}(i)$ so that the energy of excitation signal is scaled by $G^2_{ALC}(i)$. $g_p(m)$ and $g_c(m)$ remain unchanged in a sub-frame, and $G_{ALC}(i)$ is the same in a frame, so the frame index i and sub-frame index m are omitted for the sake of simplicity in the following discussion.

In [6], the modified adaptive codebook gain is calculated as:

$$g_p^{'} = g_p G_{ALC} [\sum_{n=0}^{N-1} v^2(n) / \sum_{n=0}^{N-1} (v'(n))^2]^{1/2} \tag{7}$$

where $g_p^{'}$ is the modified adaptive codebook gain, $v'(n)$ is the adaptive codebook excitation from the partial decoder, and N is the sub-frame length.

The energy of excitation signal obtained from partial decoder is as $G^2_{ALC}(i)$ times as the energy of the excitation signal obtained from the decoder, i.e.,

$$\sum_{n=0}^{N-1}(g_p^{'}v'(n)+g_c^{'}c^{'}(n))^2 = G_{ALC}^2\sum_{n=0}^{N-1}(g_p v(n)+g_c c(n))^2 \quad (8)$$

where g'_c is the modified fixed codebook gain and $c'(n)$ is the fixed codebook excitation from the partial decoder. The above equation can be written as a quadratic equation of g'_c. By solving the equation, we can get value of g'_c.

When there are probable overflow situations or the ALC gain G_{ALC} is not stationary, the fixed codebook gain modified by the method in [6] may have strong fluctuations, which will degrade the speech quality greatly at the receiver.

Therefore, we need to detect these two kinds of situations, and find a proper way for the individual modification of fixed codebook gain.

The gain stability factor R_G is used to detect the non-stationary condition in the ALC gains. Similar to the level stability factor, R_G is defined as the ratio of ALC gain in the current frame and its long-term average in the previous L_G frames, which can be expressed as:

$$R_G(i) = \frac{G_{ALC}(i)}{\frac{1}{L_G}\sum_{j=1}^{L_G} G_{ALC}(i-j)} \quad (9)$$

When R_G is larger than its threshold, the gain of ALC is considered to be non-stationary. The threshold is set to 0.9 in this paper.

When the speech amplitude after ALC exceeds the range of 16-bit quantization, the overflow will occur. In order to detect this situation, the product of ALC gain and the maximum absolute amplitude of speech in each frame is calculated. If this product is larger than 32767, the output speech of ALC will be truncated. In case of this situation, the ALC gain needs to be re-calculated as:

$$G_{ALC}(i) = MAX / \max(|x(n)|) \quad (10)$$

where MAX is set to 32767. The operator |.| is the absolute value and max (.) is the operator of maximum.

When overflow is likely to take place or the gain of ALC is not stationary, the fixed codebook gain needs to be adjusted individually. The modification rule is that, the energy of the fixed codebook excitation from the partial decoder is as $G_{ALC}^2(i)$ times as the energy of fixed codebook excitation from the decoder, which can be expressed as:

$$(g_c^{'})^2\sum_{n=0}^{N-1}(c^{'}(n))^2 = G_{ALC}^2 g_c^2 \sum_{n=0}^{N-1}(c(n))^2 \quad (11)$$

Then, the modified fixed codebook gain g'_c can be calculated as:

$$g_c^{'} = G_{ALC}\, g_c \left(\sum_{n=0}^{N-1}(c(n))^2 / \sum_{n=0}^{N-1}(c^{'}(n))^2\right)^{1/2} \quad (12)$$

The modified adaptive and fixed codebook gains are jointly-quantized to write back to the bit-stream. And the decoded speech in the receiver will be at the desired level.

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the compressed domain automatic level control method in terms of speech level bias, objective speech quality and computational complexity.

The purpose of speech level bias test and objective speech quality test is to evaluate the performance of ALC methods after the output level becomes stable. In order to eliminate the influence of the output speech with non-stationary level in the initial segments, we use the long speech signal obtained simply by coping the original speech data, and the latter half of the long output speech is extracted as the test signal for the following performance test.

The test principle is shown in Fig. 5. First, the original speech is duplicated to form the long speech signal. Then the input bit-stream is obtained by encoding the long speech signal. Next the proposed and reference ALC methods are applied to get the output bit-streams. Finally, the latter half of the decoded output speech is extracted as the test signal.

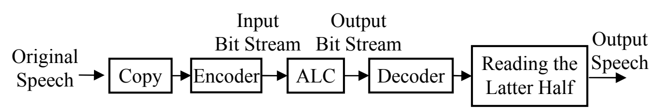


Figure 5. The block diagram for the generation of test signals

The reference method is realized on the basis of the P.56 tool. First the input bit stream is decoded, the level of decoded speech is adjusted to the target level by the P.56 tool in linear domain, and finally the scaled speech is re-encoded to obtain the output bit stream.

A. *Speech Level Bias test*

In this test, the clean speech materials are chosen from the NTT database. The noise signals, including Street, Volvo, Factory and White, are obtained from NoiseX-92 database [9]. All the test materials have been down-sampled to 16 kHz before the test. The SNR conditions of 6dB, 12dB, and 18dB are used in this test. Three bit-rates of ITU-T G.722.2 codec, including 6.6kbps, 15.85kbps, and 23.05kbps, are used in the speech level bias test. And the input speech level is set to -8dBm0, -13dBm0, and -23dBm0, respectively.

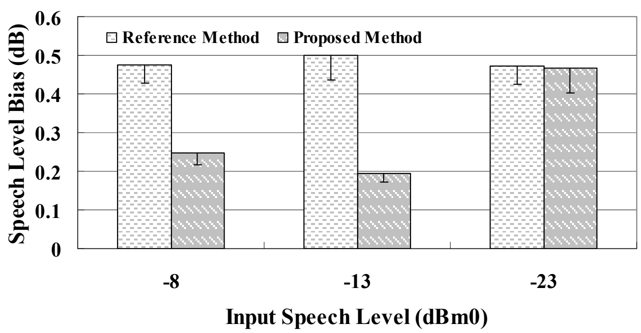
The level of the input speech is adjusted to -19dBm0 by the proposed method and the reference method, respectively. Then the absolute value of the difference between the output speech level and target level is calculated as follows:

$$\Delta = |AL_{out} - AL_{dest}| \quad (13)$$

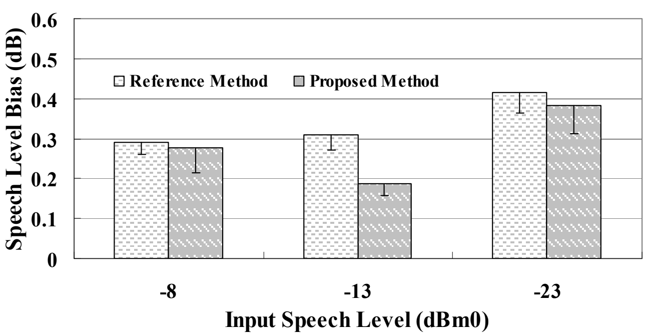
where AL_{out} is the output speech level, and AL_{dest} is the target level. AL_{dest} is set to -19dBm0 (-25.15dBov) which is known as the most comfortable speech level for the human hearing.

The results of speech level bias test in the bit-rates of 6.6kbps, 15.85kbps, and 23.05kbps are shown in Fig. 6 (a), (b), and (c), respectively. The results are averaged over different noise types and different SNR conditions. The lower boundary of 95% confidence interval has been marked on the Figure.

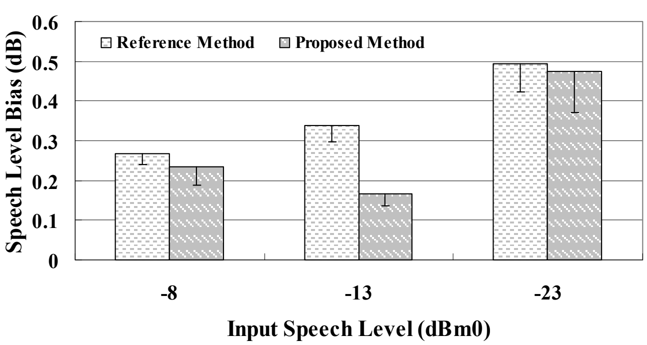
Speech level bias is the difference between the target level and the level of output speech. Fig. 6 shows that the level bias of the proposed method is less than 0.5dB in all the test conditions. The proposed method outperforms the reference method significantly at the 95% confidence level for the input speech level of -13dBm0 in all the bit-rates under test and -8dBm0 in 6.6kbps. In the other test conditions, the speech level bias of the proposed method is slightly smaller than the reference method.



(a) 6.6kbps



(b) 15.85kbps



(c) 23.05kbps

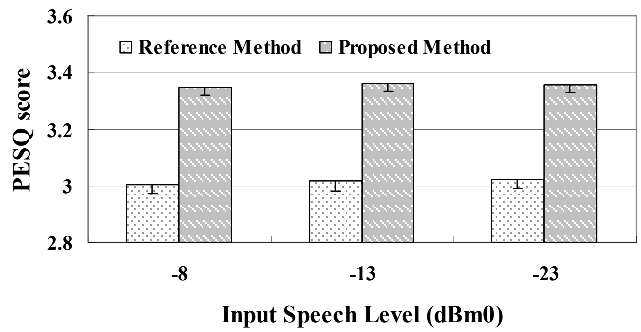
Figure 6. The results of speech level bias test at the bit-rates of (a) 6.6kbps, (b) 15.85kbps, and (c) 23.05kbps

B. *Objective Speech Quality Test*

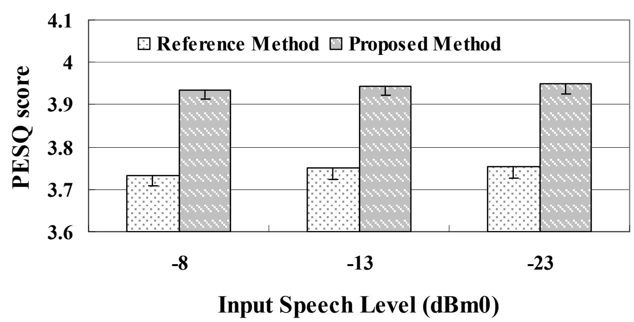
The objective speech quality of the output speech by the ALC method is evaluated by Perceptual Evaluation of Speech Quality (PESQ) score [10]. PESQ score varies from 1 to 4.5, and the higher PESQ score corresponds to better speech quality.

The test conditions are the same as the speech level bias test. The level of input speech samples is adjusted to -19dBm0 by the proposed and reference methods, respectively. The PESQ score of the output speech is obtained by comparing with the original speech.

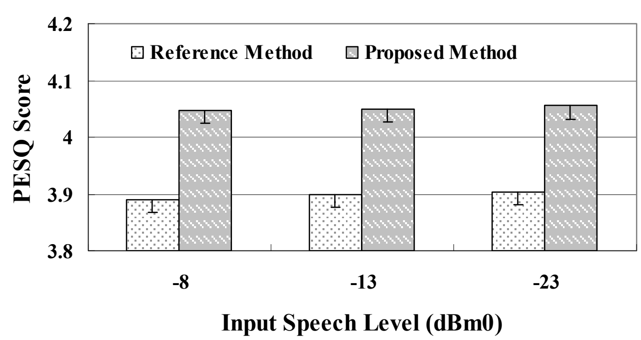
The results of objective speech quality test in the three bit-rates are shown in Fig. 7 (a), (b), and (c), respectively. The result is an average over the four noise types and the three SNR conditions. The lower boundary of 95% confidence interval has been marked on the Figure.



(a) 6.6kbps



(b) 15.85kbps



(c) 23.05kbps

Figure 7. The results of objective speech quality test at the bit-rates of (a) 6.6kbps, (b) 15.85kbps, and (c) 23.05kbps

From the test results in Fig. 7, the PESQ scores of the proposed method are 0.15 to 0.3 better than that of the reference method. The proposed method outperforms the reference method significantly at the 95% confidence level in all the test conditions. The decoding, processing and re-encoding procedure of the reference method is the main reason for the degradation of speech quality. In contrast, the proposed method just modifies the adaptive and fixed codebook gains, so the influence on the quality of speech is much lower.

C. Computional Complexity test

For the use of real-time applications, both the proposed and the reference algorithms are implemented in fixed-point C language, and the computational complexity is calculated by the tools in STL2005 under the standard of ITU-T G.191 [11]. The unit of complexity is Weighted Million Operations per Second (WMOPS). The Average and Worst Case complexities are summarized in Table Ⅰ.

TABLE I. THE COMPLEXITY IN THE FIXED-POINT C PROGRAM

Bit-rate (kbps)	**Reference Method**		**Proposed Method**	
	Average (WMOPS)	***WorstCase (WMOPS)***	***Average (WMOPS)***	***WorstCase (WMOPS)***
6.6	24.288	24.609	8.612	8.934
8.85	26.718	27.125	8.665	9.058
12.65	29.426	29.621	8.35	8.53
14.25	31.54	31.727	8.377	8.557
15.85	31.735	31.821	8.4	8.586
18.25	32.545	32.739	8.444	8.626
19.85	33.517	33.707	8.469	8.641
23.05	33.271	33.458	8.527	8.7
23.85	32.355	32.543	8.542	8.709

The results in Table I show that the proposed method saves 65% to 75% of the complexity compared with the reference method. The complexity of the reference method is concentrated in the procedure of re-encoding, while in the proposed method, only the gain parameters have to be re-quantized.

V. CONCLUSION

In this paper, a compressed domain automatic level control is proposed based on ITU-T G.722.2. The P.56 tool is adopted in the proposed method to measure the real-time level of input speech. Then the gain of ALC is calculated as the difference between the target level and the active speech level. Finally the codebook gains of G.722.2 codec are jointly modified according to the ALC gain. In the experiments, in comparison with the reference method, the proposed method has smaller speech level bias, lower computational complexity and higher objective quality. The proposed ALC method can be easily applied to the transmission network where CELP based speech codec is adopted, not only restricted to the G.722.2 codec.

ACKNOWLEDGMENT

This work was supported by Beijing Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education (KZ201110005005) and The Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality.

REFERENCES

[1] G. R. Steber, "digital signal processing in automatic gain control systems," 14th Annual Conference of Industrial Electronics Society, IECON '88, vol. 2, pp. 381-384, 1988.

[2] A. Lovrich, G. Troulinos, and R. Chirayil, "An all digital automatic gain control," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 3, pp. 1734-1737, 1988.

[3] P. L. Chu, "Voice-activated AGC for teleconferencing," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 929-932, 1996.

[4] M. Mohammed and K. E. Bijoy, "An intelligent automatic level controller for speech signals: BELBIC," Signal Processing, Communications and Computing (ICSPCC), pp. 1-5, 2011.

[5] ITU-T Rec. P.56, Objective measurement of active speech level.

[6] R. A. Sukkar, R. Younce, and P. Zhang, "Dynamic scaling of encoded speech through the direct modification of coded parameters," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. I-677~I-680, 2006.

[7] ITU-T Rec. G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB).

[8] ITU-T Rec. G.100.1, The use of the decibel and of relative levels in speechband telecommunications.

[9] A. Varga and J. M. SteenekenH, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, pp. 247-251. 1993.

[10] ITU-T Rec. P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[11] ITU-T Rec. G.191, Software tools for speech and audio coding standardization.